ChatGPT 生成与学者撰写文献摘要的对比研究——以信息资源管理领域为例

张强[1]，王潇冉[2]，高颖[1]，周洪[3,4]
（1. 华中师范大学信息管理学院 武汉 430079；2.安徽工程大学计算机与信息学院 芜湖 241000；3.中国科学院大学经济与管理学院 北京 101190；4.中国科学院武汉文献情报中心 武汉 430071）

摘要：[目的/意义]探究 ChatGPT 生成与学者撰写的中文论文摘要之间的异同，并分析二者之间的内容特征差异，为 AI 生成学术论文检测及相关研究提供借鉴。[方法/过程]首先，以信息资源管理领域为例，分别抽取了图书馆学、情报学、档案学近三年各 500 篇高被引论文，基于获取的论文题目采用 Prompt 方式应用 ChatGPT 工具生成对应的摘要文本，构建数据集合；其次，采用了 9 种机器学习及深度学习算法对 ChatGPT 生成与学者撰写的摘要文本进行分类检测；最后，从文本特征、主题模型、ROUGE 评测对二者的异同进行多角度分析，从而揭示二者之间的异同点。[结果/结论]基于数据集所训练的主流机器学习及深度学习算法可以有效地分辨摘要是 AI 生成还是学者撰写，其中 BERT 和 ERNIE 的效果最好，而机器学习算法中 RF 和 Xgboost 效果最好。ChatGPT 生成的摘要字符数量、句子数量较学者撰写的要多，关键词多为模版化的转折性词语；两者的文本主题大部分相同，在“学科体系”、“数字人文”等主题上存在差异；ROUGE 及余弦相似度定量分析表明 ChatGPT 生成的摘要与学者撰写的摘要文本存在明显的“形似”而非“神似”的现象。

关键词：ChatGPT　文本分类　文本特征　论文摘要
分类号：G353

2022 年末，ChatGPT 一经面世就成为史上用户增长速率最快的消费级应用，标志着生成式人工智能（Artificial Intelligence Generated Content, AIGC）成为学界与业界新的研究热点[1]。所谓生成式人工智能就是一种通过在大规模语料库中学习所生成新的数据、文本、图像等内容的新一代人工智能，其在自然语言处理、图像生成、机器翻译、语音生成、艺术创作等领域均具有广泛的应用场景，有望引发新一轮的科技革命与产业重构[2]。

ChatGPT 作为一款聊天式的交互对话应用，代表了当前 AIGC 产业化的最高水平，具备极强的自然语言理解与生成能力。可以根据用户的提示信息（Prompt）来理解用户意图并生成相应的答案[3]。在教育领域，有关调查显示，美国有 89%的大学生正使用 ChatGPT 来完成学术作业，这表明 ChatGPT 已具备初级科研工作者的水准，生成的学术论文具有格式完整、逻辑流畅等特征[4]。在学术领域，有学者将其署名为合作作者，由此引发了关于 ChatGPT 生成内容著作权的归属问题纠纷。Nature 针对 ChatGPT 被列为作者等问题在投稿指南中新增了大语言模型不能列为论文作者和论文中使用了大语言模型需要在方法或致谢部分进行明确说明两大原则[5]。国内以《图书情报工作》为代表的学术期刊也在投稿政策说明中明确声明不接受署名包括 AI 工具的学术论文投稿等原则[6]。由此可见，学术期刊对 ChatGPT 所引发的学术伦理问题的高度重视，亟需可以分辨学术论文是否由 AI 生成的判定方法与标准。

基于上述分析，甄别学术论文内容是否由 ChatGPT 类 AI 工具生成以及所生成的文本内容与学者人工撰写的特征差异就显得尤为重要。具体来说，本文以信息资源管理学科下的图书馆学、情报学、档案学领域的中文学术论文摘要为研究对象，主要研究如下问题：（1）统计机器学习及深度学习方法能否判别出中文学术论文摘要是由学者撰写还是 AI 生成？（2）学者撰写的学术论文摘要与 AI 生成的学术论文摘要在文本特征上具有哪些异同？本文研究可以为 AI 生成学术论文文本的质量评价提供参考，有助于期刊对学术论文的原创性进行辅助评判。同时，根据人工与 AI 生成中文学术论文的摘要进行内容特征分析，探究 AI 生成

内容的特征、质量及与人工对比的优劣之处，从而推动 AI 工具在学术论文撰写、学术出版伦理等方面的合理使用。

1 相关研究

ChatGPT 的官方网站 OpenAI 上介绍了 ChatGPT 模型背后的方法，其背后的核心技术包括基于 Transformer 的预训练模型、人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）、监督微调训练、奖励模型[7]。简而言之，ChatGPT 是在预训练之后通过监督微调、奖励模型与强化学习等技术手段来进一步优化模型从而生成合理、流畅的对话信息，并使 ChatGPT 具有人类的常识与价值观，对待敏感问题会进行合理的规避。当前，针对 ChatGPT 的相关研究主要包括如下两方面：

一是关于 ChatGPT 对某一学科或领域发展的冲击与影响。Chris 与 Richard[8]认为 ChatGPT 类生成式人工智能技术可以加速科学研究、生成创新性假设，从而推动知识的发展，但对数据偏见、文本伦理、科学研究的重复性等方面表示了担忧。Pawan 等人提出了生成式人工智能介入人力资源管理领域学术研究的发展路径，将其与人力资源管理过程、实践、关系和结果等各个方面联系起来，探析了未来人力资源管理研究的方向[9]。戴岭等人认为 ChatGPT 类人工智能技术突破了时空与个体间的障碍，串联了学习网络中古今中外的各个领域，有利于教育行业的数字化转型和教育生态系统的变革，但是也为教育伦理和教育数据安全带来了挑战[10]。在信息资源管理领域，主要有陆伟等人从支撑算法与技术、信息资源建设、信息组织与信息检索、内容安全与评价、人机智能交互与协同六个方面探讨了以 ChatGPT 为代表的大语言模型对信息资源管理的影响[11]。张智雄等人通过总结生成式人工智能的发展历程，从数据组织方式、知识服务模式、情报分析方法、文献使用方式、文献情报队伍建设等方面分析了 ChatGPT 对文献情报工作的影响，并根据文献情报工作的特点给出了 ChatGPT 时代下文献情报工作的发展建议，认为知识获取能力的提升是生成式人工智能技术高速发展的本质，高价值的语料库是生成式人工智能的基础，文献情报领域管理着蕴含人类高价值知识的领域，在生成式人工智能时代需要主动适应和发展[12]。Brady D.等人通过概述了 ChatGPT 作为一个聊天机器人背后的技术原理，接着利用访谈讨论了 ChatGPT 在图书馆领域的搜索与发现、参考与信息服务、编目与元数据生成、内容创建等方面大有可为，但是仍需要警惕隐私与偏见等伦理问题[13]。曹树金等人从研究问题、研究数据和研究范式三个角度探究了生成式人工智能对情报学研究的影响，并从四个服务层面来分析生成式人工智能对情报实践工作的变化，认为情报学在保证客观审视的态度基础上积极拥抱新一代人工智能[14]。周文欢[15]通过分析档案领域的数字化与智能化研究现状的基础上，分析了 ChatGPT 可以在档案文本摘要、档案分类、档案信息智能检索、档案信息知识问答和档案保护和安全五个方面具有广阔的应用前景，可提高档案管理的效率、精度和智能化水平。

二是关于 ChatGPT 在各类文本生成任务中的表现与测评。2023 年 OpenAI 公司发布了基于 GPT-4 的 ChatGPT 在各类考试任务中的表现，在美国律师资格考试中分数超过了 90%的人类。Zheng 等人考虑到 ChatGPT 的训练数据集来源于 2021 年之前，基于一篇不存在 ChatGPT 数据库中的学术论文反复向 ChatGPT 进行提问从而评估其表现，结果表明当前的 ChatGPT 还无法胜任科学写作任务，但对于检查语法错误并改进语言具有益处[16]。Fredricton[17]认为在使用 ChatGPT 用于写作前需要知道 ChatGPT 会编造不存在的引文内容，作者不该在科学写作中使用 ChatGPT 生成的内容。他还认为期刊没有必要对作者要求其表明在科学写作中使用 ChatGPT 所起的作用，ChatGPT 与词库、语法检查器等一样都是写作工具，ChatGPT 只是作者所选择的工具之一，作者本人需要对自己的决定负责。在中文领域，张华平等人将 ChatGPT 与多个已有的预训练模型进行对比，发现 ChatGPT 在中文情感分析任务上已经具有较高的准确率，但是在中文领域的闭卷问答上会经常出现事实性的错误[18]。鲍彤等人则是在多个中文公开数据集上将 ChatGPT 与多个预训练模型对比分析其在实体抽

取、关系抽取和事件抽取上的效果，结果表明 ChatGPT 在事件抽取任务上的表现优于其它两类任务[19]。施亦龙等人从外部与内部特征、情感与认识等方面探讨了同一问题下 ChatGPT 与知乎人工高赞答案之间的优劣特征。发现 ChatGPT 可以使人更加便捷地获取想要的信息，回答的文本特征接近人工高赞的答复，但是在不同主题下回答的质量差异较大，并伴随着虚假信息[20]。

综上所言，除了探讨 ChatGPT 类生成式人工智能工具给学科或领域带来的机遇与挑战之外，也开始有相关的实证研究来对其表现进行测评，但就中文语料库的相关研究还相对不足。针对中文期刊论文，分析 ChatGPT 类生成式人工智能工具生成的文本内容与学者人工撰写的学术内容之间异同的相关研究还远远不足。本文选择信息资源管理领域下的图书馆学、情报学、档案学的学术文献摘要文本作为基础数据集，利用 ChatGPT 根据文献题目来生成对应的摘要文本，探究 ChatGPT 在中文学术论文生成上的性能表现，并分析二者之间的差异。

2 研究设计

为了分析学者人工撰写的学术论文摘要和 ChatGPT 生成的摘要文本之间的异同，本文设计的研究框架如图 1 所示。

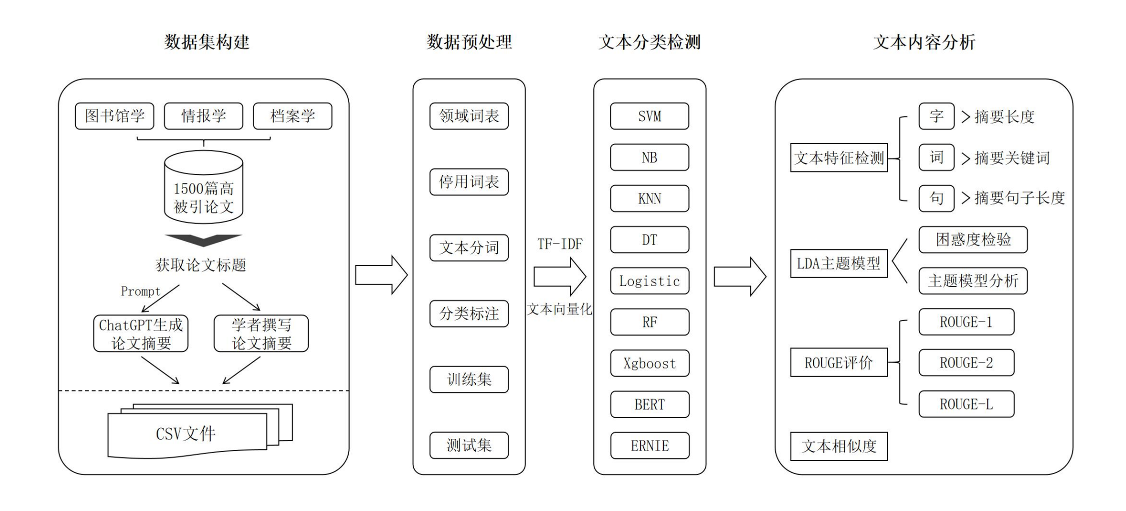


图 1 研究框架

2.1 数据来源与处理

本文以信息资源管理领域的期刊论文为研究对象，并依据二级学科划分为图书馆学、情报学、档案学三类，考虑到学术文献的代表性，本文选择了核心期刊作为文献来源。同时考虑到 ChatGPT 的训练数据更新时间为 2021 年 9 月（目前，GPT4 的训练数据截止时间也是 2021 年 9 月），最终确定三类学科分别筛选出 2018 年 9 月-2021 年 8 月间 500 篇高被引文献共计 1500 篇作为基础研究样本，针对跨领域的期刊如《图书情报工作》则人工介入进行分类筛选，具体的来源期刊名称及论文数量如表 1 所示。

表 1 来源期刊名称及论文数量

二级学科名称	来源期刊名称	论文数量
图书馆学	《中国图书馆学报》	100
	《图书情报工作》	100
	《图书情报知识》	100
	《大学图书馆学报》	100
	《国家图书馆学刊》	100
情报学	《情报学报》	100

	《图书情报工作》	100
	《图书情报知识》	100
	《情报资料工作》	100
	《情报理论与实践》	100
档案学	《档案学研究》	250
	《档案学通讯》	250

在获取到相关论文的题目之后，还需要 ChatGPT 来生成对应的摘要文本。在生成式 AI 工具的使用过程中 Prompt 的重要性不言而喻，Prompt 是一种包含了引导性提示信息的语言，从而让模型更好地理解并生成内容。可以说 Prompt 的好坏直接影响到模型的输出结果。本文参考了 CRISPE 框架来撰写 Prompt，CRISPE 框架将提示的创建过程拆分为清晰、结构化的步骤[21]。其中，CR（Capacity and Role）代表能力与角色，即提问者希望 ChatGPT 扮演何种角色。I（Insight）为 ChatGPT 提供背景信息与上下文，以便让 ChatGPT 充分了解背景与需求。S（Statement）代表用户所制定的明确的任务目标，以便 ChatGPT 满足用户的回应。P（Personality）代表用户希望 ChatGPT 以何种风格来进行回应，这一步骤有助于 ChatGPT 生成的内容具备个性化。E（Experiment）代表用户在粗略搜索的情况下要求 ChatGPT 生成多种示例，生成多种答案，从而让用户可以在多样选择中进行对比和评估。本文最终确定的要求 ChatGPT 生成学术论文摘要的 Prompt 为：

“假设你是一位从事图书馆学/情报学/档案学的科研工作者，你已经构思了一篇新的学术论文，题目为“XXX”，请你以一名中国科研工作者的身份及中文学术期刊的要求来撰写该题目的摘要文本。”

由于 GPT-4 目前有调用和问答次数限制，本文通过自编 Python 代码调用 GPT3.5 接口来批量获取 ChatGPT 生成的摘要内容，调用代码如图 2 所示。

```
import openai
import time
1 个用法
def generate_prompt(prompt):
    return "假设你是一位从事图书馆学/情报学/档案学的科研工作者，你已经构思了一篇新的学术论文，题目为“XXX”，请你以一名中国科研工作者的身份及中文学术期刊的要求来撰写该题目的摘要文本。" + prompt
1 个用法
def openai_reply(content, apikey):
    openai.api_key = apikey
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo-0301",
        messages=[
            {"role": "user", "content": generate_prompt(content)}
        ],
        temperature=0.5,
        max_tokens=1000,
        top_p=1,
        frequency_penalty=0,
        presence_penalty=0,
    )
    return response.choices[0].message.content
# 从文件中读取题目列表
1 个用法
def read_topics_from_file(file_path):
    topics = []
    with open(file_path, 'r', encoding='utf-8') as f:
        for line in f:
            topic = line.strip()
            if topic:
                topics.append(topic)
    return topics
# 从 topics.txt 文件读取题目列表
file_path = 'topics.txt'
topics_list = read_topics_from_file(file_path)
# 使用 OpenAI 逐一生成题目的摘要，并将摘要保存到文件
output_file = 'summaries.txt'
with open(output_file, 'w', encoding='utf-8') as f:
    for i, topic in enumerate(topics_list, 1):
        print(f"生成摘要 {i}/{len(topics_list)}: ")
        response = openai_reply(topic, "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")
        f.write(f"摘要 {i}: \n{response}\n")
        print("------------------------")
        # 添加延迟，等待一段时间再发送下一个请求
        time.sleep(20)
```

图 2 调用 ChatGPT 生成学术文本摘要代码

最终，将学者撰写的 1500 篇摘要文本与 ChatGPT 生成的 1500 篇摘要文本保存为本地

Excel 文件。数据预处理的过程如下所述：

（1）领域词表构建：将学者撰写的 1500 篇论文关键词作为初始领域词表，加上网络中图情档领域的常规术语，经人工筛查后共确定 1376 个词语作为领域词表，以便后续的分词操作。

（2）停用词表构建：为了尽可能获取 ChatGPT 生成的摘要文本特征，本文在停用词的选择上仅考虑将标点符号和无意义虚词加入到停用词表中。

（3）文本分词：通过自编 Python 代码调用 LTP 自然语言处理包，加载领域词表和停用词表来对摘要文本进行分词。

（4）分类标注：对 ChatGPT 生成的与学者撰写的文本摘要分别以 0 和 1 进行标记。

2.2 文本分类标注

本文的研究目标是探究当前的主流机器学习和深度学习算法能否鉴别出 ChatGPT 生成和学者撰写学术论文摘要的类别及其差异。从粗粒度来看可将这一问题转化为经典的二分类问题，通过 TF-IDF 法来进行文本向量化表示后，利用 SVM、NB、K 近邻、决策树、逻辑回归、随机森林、xgboost 常见的七种机器学习分类算法及 BERT 和 ERNIE 两种深度预训练语言模型进行分类实验，选取准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值作为评价指标，按照常规机器学习领域的数据集划分标准，将文本数据集的 70%作为训练集来训练分类模型，剩余的 30%数据集作为测试集来评估模型的分类性能。

2.3 文本内容分析

除了对 ChatGPT 生成与学者撰写的学术论文摘要做分类识别之外，本文还采用了文本特征检测、主题模型一致性检测、ROUGE 评测来从内容层面对二者之间的差异进行解读。

文本特征检测主要包括字、词、句三个维度的特征检测。具体而言，字就是判断 ChatGPT 生成与学者撰写的摘要字数差异；词是二者之间高频关键词的异同；句是二者之间摘要句子个数的差异。文本特征通常可以直观反映一个文本的核心特点，通过从字、词、句角度来分析有利于比较二者在重点概念、学术术语、语言表达上的差异。

主题模型一致性检测主要利用 LDA 主题模型来对比二者之间的主题分布。LDA（Latent Dirichlet Allocation）模型由 BLEI 等学者于 2003 年提出，该模型是一种包括词、主题和文档的三层贝叶斯网络模型，用于发现文本数据中的潜在主题并将文本文档分配到这些主题当中。使用 LDA 主题模型可以识别 ChatGPT 生成与学者撰写的学术论文摘要之间的主题差异，从而揭示出它们在内容上的异同。LDA 主题模型中文档生成的过程如图 3 所示。

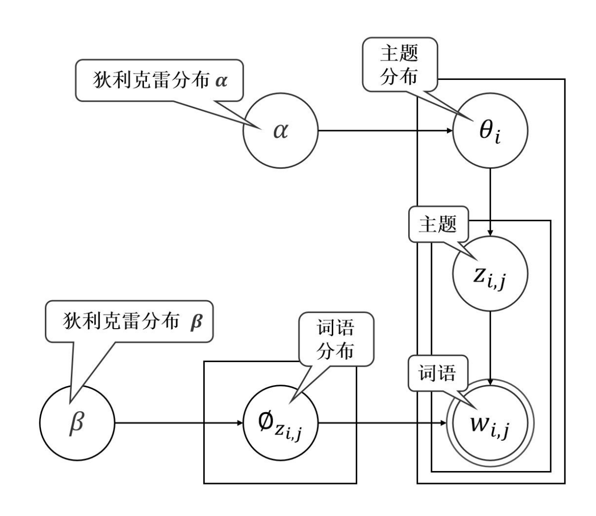


图 3 LDA 主题模型图

第一步：按照先验概率 $p(d_i)$ 选择一篇文档 d_i；

第二步：从 Dirichlet 分布 α 中，抽样生成文档 d_i 的主题分布 θ_i；

第三步：从主题分 θ_i 中，抽样生成文档 d_i 的第 j 个词的主题 $z_{i,j}$；

第四步：从 Dirichlet 分布 β 中，抽样生成主题 $z_{i,j}$ 对应的词语分布 $\phi_{z_{i,j}}$；

第五步：从词语分布 $\phi_{z_{i,j}}$ 中，抽样生成词语 $w_{i,j}$。

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是一组自动评价指标，用来衡量生成的文本摘要或机器翻译结果与参考摘要之间的相似度。在自然语言处理任务中评估摘要生成和机器翻译的任务重被广泛使用。ROUGE 主要关注召回率（Recall），将生成摘要和参考摘要看作是一个词袋模型，通过计算词的重叠程度来衡量它们之间的相似性，即判定生成的摘要中包含了多少参考摘要的内容。常见的 ROUGE 指标主要包括：

ROUGE-N：该指标计算生成摘要和参考摘要之间 N-gram（连续 N 个词）的召回率。计算方式如公式 1 所示。

$$ROUGE-N=\frac{\sum_{S\in\{referenceSummaries\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{referenceSummaries\}}\sum_{gram_n\in S}Count(gram_n)} \quad \text{公式（1）}$$

其中，分母是学者撰写摘要中 $n-gram$ 的个数，分子是学者撰写摘要与 ChatGPT 生成的摘要共现的 $n-gram$ 的个数，$ROUGE-N$ 特点是简洁且有词序特征，但是随着 N 的增大，值会骤降。一般采用 $ROUGE-1$ 和 $ROUGE-2$ 作为评价指标。

ROUGE-L：该指标计算最长公共子序列（Longest Common Subsequence，LCS）的召回率。它衡量了生成摘要和参考摘要之间的长距离依赖和顺序一致性，计算方式如公式 2-4 所示。

$$R_{lcs}=\frac{LCS(X,Y)}{m} \quad \text{公式（2）}$$

$$P_{lcs}=\frac{LCS(X,Y)}{n} \quad \text{公式（3）}$$

$$F_{lcs}=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}} \quad \text{公式（4）}$$

其中，$LCS(X,Y)$ 是 X 和 Y 最长公共子序列的长度，m 与 n 分别代表学者撰写的摘要与 ChatGPT 生成摘要的长度，R_{lcs} 与 P_{lcs} 分别代表召回率和准确率，β 用于均衡二者之间的重要性。$ROUGE-L$ 的特点是不需要像 $ROUGE-N$ 去制定 $n-gram$ 的长度，但是只考虑了最长子序列的长度，比较适合短摘要提取的测评。

此外，在该步骤本文还采用了余弦相似度来检测 ChatGPT 生成与学者撰写摘要的相似程度，从而跟 ROUGE 的指标结果形成对比。

3 结果分析

3.1 摘要分类结果

基于前文所述的二分类步骤及 9 种分类模型，本文对 ChatGPT 生成的学术论文摘要和学者人工撰写的摘要进行分类测试，结果如表 2 所示。

表 2 ChatGPT 生成与学者人工撰写文本摘要在不同分类模型下的对比效果

分类模型	评价指标：Accuracy(A) Precision(P) F1-Score(F1)											
	图书馆学			情报学			档案学			整体		
	A	P	F1	A	P	F1	A	P	F1	A	P	F1
SVM	90.33%	90.61%	90.29%	91.00%	91.06%	91.00%	79.67%	80.10%	79.67%	94.12%	94.18%	94.11%
NB	63.33%	69.22%	60.96%	73.33%	75.53%	72.57%	69.00%	70.10%	67.84%	75.36%	78.18%	74.89%
KNN	45.67%	45.11%	42.44%	46.67%	46.00%	45.27%	49.00%	47.21%	45.79%	47.84%	47.98%	47.61%
DT	92.67%	92.67%	92.67%	91.33%	91.37%	91.33%	88.67%	88.61%	88.64%	93.12%	93.11%	93.12%
Logistic	91.33%	92.04%	91.27%	87.00%	87.23%	86.99%	75.00%	75.73%	75.00%	92.45%	92.76%	92.43%
RF	94.67%	94.66%	94.66%	96.33%	96.55%	96.32%	95.00%	95.11%	95.00%	96.12%	96.11%	96.12%
Xgboost	94.00%	93.99%	94.00%	97.00%	97.00%	97.00%	93.67%	93.62%	93.65%	96.56%	96.55%	96.56%
BERT	99.67%	99.68%	99.67%	**100%**	**100%**	**100%**	98.67%	98.75%	98.66%	97.89%	97.91%	97.89%
ERNIE	**100%**	**100%**	**100%**	99.67%	99.65%	99.67%	**99.00%**	**98.98%**	**99.00%**	**99.45%**	**99.45%**	**99.45%**

由表 2 可知，在 9 种分类模型中，基于深度学习的 ERNIE 取得的分类效果最好，BERT 模型次之。原因在于本文所研究对象为中文学术论文摘要，而 ERNIE 是百度在基于 BERT 模型基础上针对中文 NLP 任务做的进一步优化。此外，在 7 种机器学习分类模型中，除了 NB 和 KNN 以外，其余 5 种机器学习分类模型的整体 F1-Score 均超过了 90%，这表明它们均具备良好的分类效果，而 NB 和 KNN 在该问题上的效果较差。从三个二级学科来看，各分类算法在档案学领域的分类上较图书馆学和情报学更低，表明从文本分类角度来看，档案学领域 ChatGPT 生成与学者撰写的摘要更加接近。

文本分类实验中，其特征词对于分类的判定至关重要，特征词的选择直接关系到模型分类的效果与能力。本文选择了在三个二级学科领域 F1-Score 均超过 90%的机器学习分类模型 RF 和 Xgboost，分析 2 种分类算法排名前 10 的特征词，结果如表 3 所示。

表 3 RF 算法与 Xgboost 算法前 10 特征词

序号	RF 特征词			Xgboost 特征词		
	图书馆学	情报学	档案学	图书馆学	情报学	档案学
1	本文	本文	探讨	本文	结论	可靠性
2	最后	结论	意义	访谈	情境	探讨
3	探讨	最后	提出	因素	图谱	调整
4	提出	探讨	研究成果	阶段	计算	实验
5	研究成果	意义	数字化	社交	典型	探究
6	结论	研究成果	重要性	探讨	生成	结果表明
7	数字化	一种	保护	最后	结果显示	比较
8	现状	提出	阐述	参考价值	意义	有利于
9	挑战	参考价值	建议	公共卫生	得出	网络
10	建议	发现	参考	实验	关键因素	研究成果

通过表 3 不难发现，尽管二种算法在三个领域的关键词及排序各有不同，但如“本文”、“结论”、“探讨”、“最后”等词在两类算法中均有出现，表明这些词可以有效的区分摘要是由 ChatGPT 生成还是学者撰写的。更具体地，RF 与 Xgboost 相比在图书馆学领域 F1-Score 较为接近，而在情报学领域 Xgboost 优于 RF，在档案学领域 RF 优于 Xgboost。结合特征词来看，表明在情报学领域“情境”、“图谱”“计算”等词语在分类上具有更重要

的特征，在档案学领域“意义”、“数字化”、“重要性”、“保护”等词语在分类上具有更重要的特征。

3.2 文本内容分析

除了验证机器学习模型是否可以区分学术论文摘要是由 ChatGPT 生成还是学者撰写以外，还需要从文本内容层面来检验二者的异同，以了解其文本内部的差异性。本节从文本特征、主题模型、ROUGE 检测三方面来探析二者之间的差异。

3.2.1 文本特征分析

学术论文摘要作为一种典型的短文本，其字、词、句均能反应摘要文本的特征。摘要长度指的是一篇学术论文摘要的中文字符个数，分别对二者的摘要长度按照三个二级学科进行统计并绘制统计直方图，如图 4 所示。

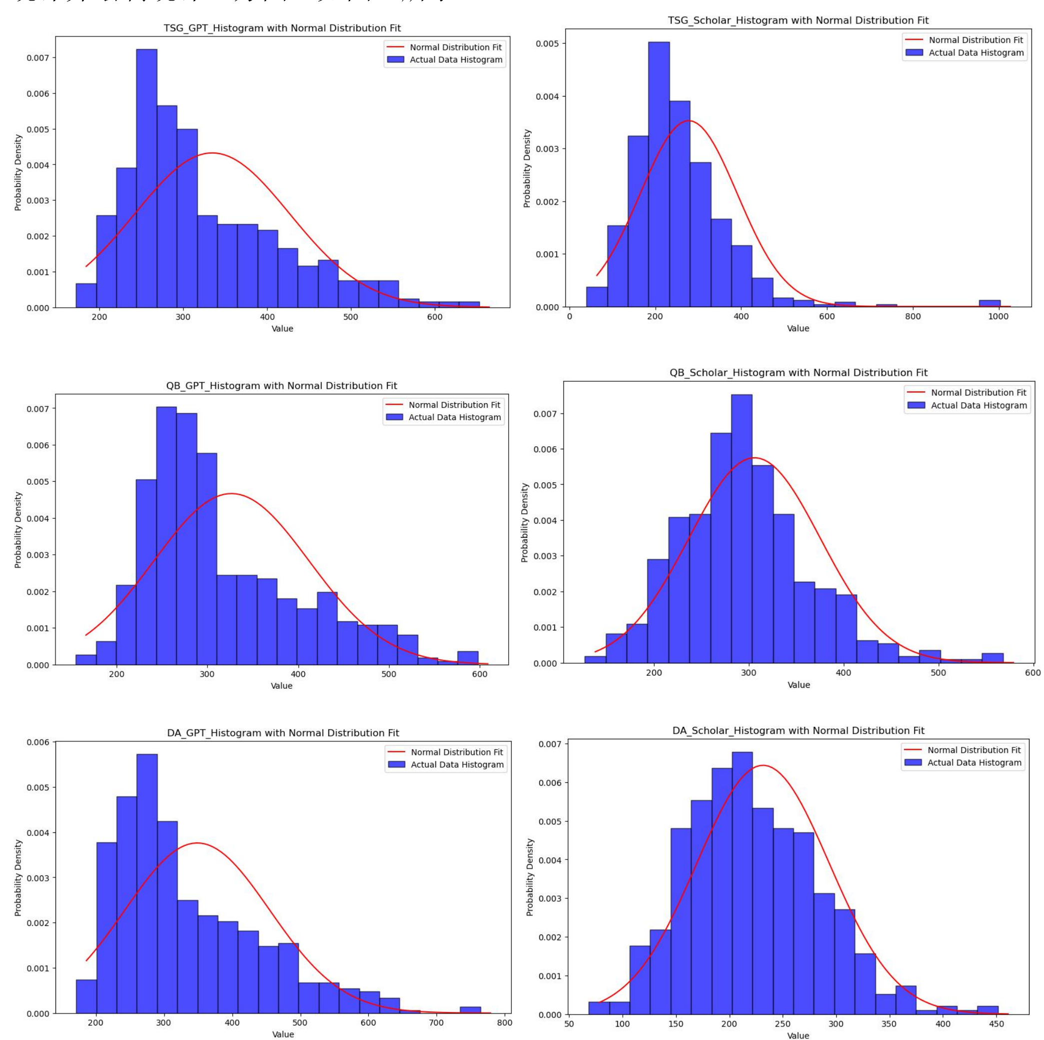


图 4 摘要长度正态分布拟合直方图

为了直观展示 ChatGPT 生成与学者撰写摘要长度的统计信息，本文采用了平均值、均方差、偏态、峰度作为统计指标，具体如表 4 所示。

表 4 摘要长度统计信息表

统计指标	图书馆学		情报学		档案学		整体	
	GPT	学者	GPT	学者	GPT	学者	GPT	学者
平均值	334.45	277.95	326.44	306.02	348.77	231.56	336.55	271.84
均方差	92.29	113.11	85.49	69.41	106.05	61.98	95.44	89.96
偏态	1.02	2.25	0.97	0.72	1.08	0.47	1.09	1.76

峰度	0.52	11.10	0.29	1.16	0.83	0.33	0.86	10.36

结合图 4 和表 4 可知，从整体上来看 ChatGPT 生成摘要文本的字符数均值为 336.55，而学者撰写摘要文本的字符数平均值为 271.84，两者之间差异性较大，但在均方差上差异较小，表明二者的数据分散程度相当。峰度差距很大，表明二者的峰值尖锐程度差距较大，学者撰写摘要文本字符数分布的峰值更加尖锐。具体到二级学科来看，在平均值指标上，情报学差距最小，档案学差距最大；在均方差指标上，情报学差距最小，档案学差距最大；在偏态指标上，情报学差距最小，图书馆学差距最大；在峰度指标上，情报学差距最小，图书馆学差距最大。

除了从字的层面来看二者之间的差异外，通过用词的习惯也能反应二者的写作风格。通过 TF-IDF 和 TextRank 算法来对二者进行关键词抽取，结果如表 5 所示。

表 5 ChatGPT 生成与学者撰写摘要文本关键词

学科名称	TF-IDF		TextRank	
	GPT	学者	GPT	学者
图书馆学	图书馆、本文、阅读、探讨、数据、提出、数字、建设、最后、影响	图书馆、数据、阅读、建设、数字、影响、智慧、提出、文献、素养	本文、图书馆、探讨、数据、阅读、最后、提出、建设、影响、数字	图书馆、数据、阅读、建设、数字、提出、文献、意义、影响、文章
情报学	本文、数据、网络、影响、提出、舆情、探讨、因素、社交、最后	数据、网络、影响、舆情、意义、结论、因素、提出、事件、治理	本文、探讨、数据、因素、影响、网络、提出、社交、最后、发现	数据、意义、结论、网络、影响、舆情、提出、本文、特征、治理
档案学	本文、探讨、数字、档案学、提出、档案管理、数字化、数据、最后、保护	数据、数字、治理、档案学、建设、文件、电子、提出、记忆、档案馆	本文、探讨、档案学、最后、档案管理、提出、数字化、数据、数字、保护	数据、治理、建设、档案学、数字、提出、记忆、本文、新、档案管理
整体	本文、图书馆、数据、探讨、提出、最后、影响、数字、数字化、建设	数据、图书馆、影响、网络、数字、意义、提出、建设、治理、结论	本文、探讨、图书馆、数据、最后、提出、影响、因素、数字、发现	图书馆、数据、治理、建设、意义、提出、影响、数字、本文、网络

由表 5 可知，从整体上看二者的关键词较为接近，但 ChatGPT 生成摘要文本往往有“本文”、“提出”、“最后”等过渡性结构化词语，而学者撰写的摘要文本则多为实质性的名词或动词。从三个二级学科来看，情报学领域二者之间的关键词较为接近，图书馆学和档案学二者间的关键词差异更多。

在从字、词角度考虑外，本文还从句子数量的角度来分析 ChatGPT 生成与学者撰写摘要的差异，力求从多个角度剖析二者的异同。分别对二者的摘要句子长度按照三个二级学科进行统计并绘制统计直方图，如图 5 所示。

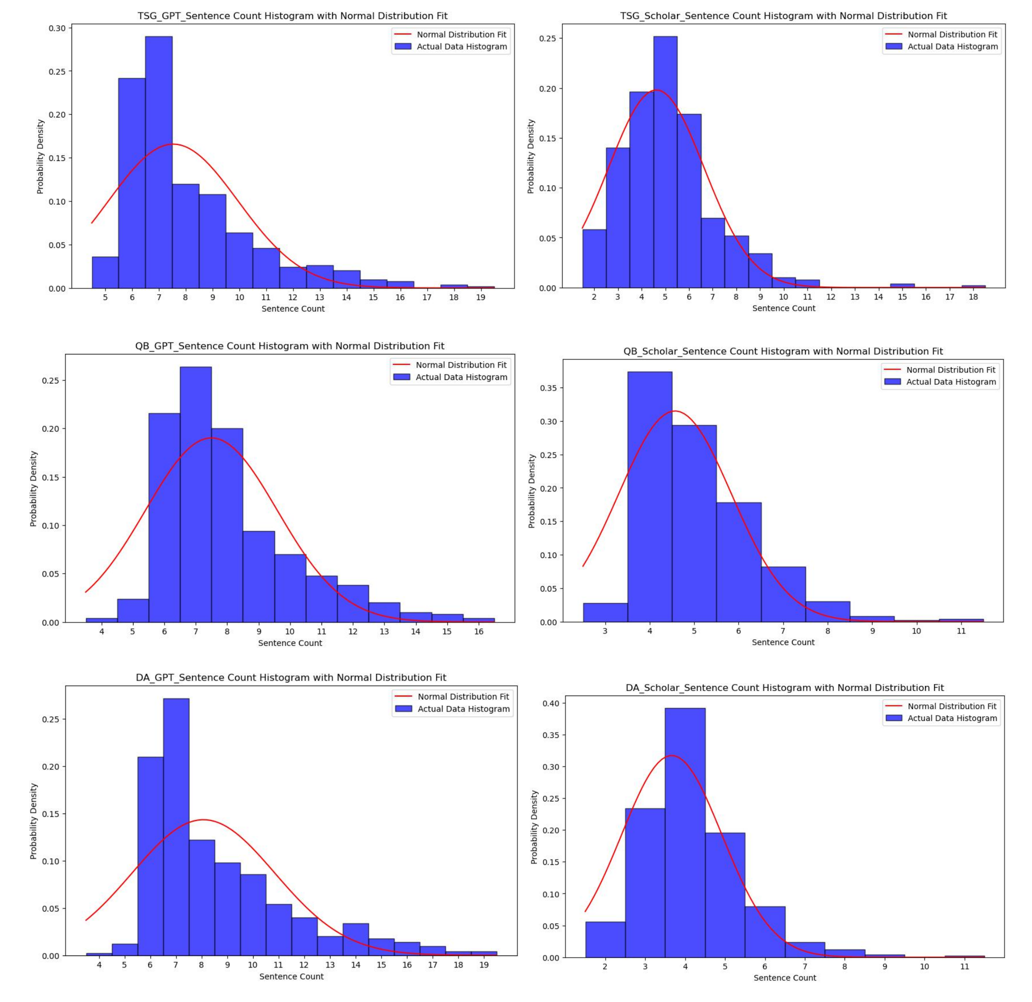


图 5 摘要句子长度正态分布拟合直方图

与分析摘要长度的统计信息类似，为了更加直观展示 ChatGPT 生成与学者撰写摘要句子长度的统计信息，这里延续采用了平均值、均方差、偏态、峰度作为统计指标，具体如表 6 所示。

表 6 摘要长度统计信息表

统计指标	图书馆学		情报学		档案学		整体	
	GPT	学者	GPT	学者	GPT	学者	GPT	学者
平均值	8.04	5.12	8.00	5.07	8.57	4.17	8.20	4.79
均方差	2.41	2.01	2.10	1.27	2.78	1.26	2.46	1.61
偏态	1.58	1.52	1.28	1.27	1.37	1.16	1.49	1.55
峰度	2.80	6.01	1.78	2.69	1.54	3.60	2.35	6.40

结合图 5 和表 6 可知，整体而言，ChatGPT 生成摘要是学者撰写摘要的文本句子数量的近 2 倍，均方差也更高，表明 ChatGPT 生成摘要的句子数量整体上更加分散。二者的偏态相当，即二者数量分布的态势较为一致。而学者撰写的摘要句子数量的峰度更大，表明其数量分布的峰值较为尖锐。具体到各二级学科来看，ChatGPT 生成摘要的句子数量均超过学者撰写摘要的句子数量；均方差方面，图书馆学领域相当，情报学与档案学均是学者撰写摘要的句子数量均方差更低；偏态方面，图书馆学和情报学两者都较为接近，而档案学则是学者撰写摘要的句子数量偏态更低；峰度方面，图书馆学与档案学领域的 ChatGPT 与学者撰写摘要的句子数量峰度差距更大，且学者撰写摘要句子数量的峰度值均超过了 3，代表这两个领域句子数量分布的峰值较为尖锐。

3.2.2 主题模型分析

对 ChatGPT 生成与学者撰写的学术论文摘要进行 LDA 主题模型分析，从而把握二者之间在文本主题上的差异。本文采用困惑度来衡量模型，具体如图 6 所示。困惑度用来描述主题的相似性，一般而言困惑度的值越低越好，但是当主题数目过多时，模型往往已经过拟合。根据所绘制的 ChatGPT 生成与学者撰写摘要文本的困惑度-主题数折线图，综合考虑二者的可能最优主题数，本文最终确定 9 为最佳主题数。

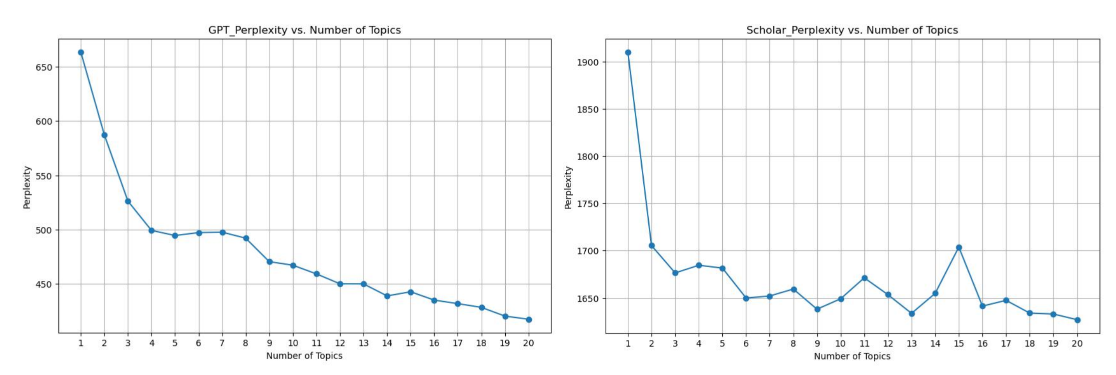


图 6 ChatGPT 生成与学者撰写摘要困惑度-主题数折线图

在确定了最优主题数之后，将经过预处理的文本导入 LDA 主题模型训练，得到“主题-词”分布，选取 9 个主题中概率排名前 5 位的词汇进行汇总，如表 7 所示。

表 7 “主题-词”分布

序号	主题名称(学者)	主题词(学者)	主题名称(ChatGPT)	主题词(ChatGPT)
1	红色记忆	疫情、红色、记忆、文献、建设	网络舆情	本文、网络、舆情、疫情、情感
2	网络舆情	网络、舆情、事件、情感、意义	智慧图书馆	图书馆、本文、建设、阅读、探讨
3	学科体系	学科、情报、应急、情报学、图书	数字人文	数字、本文、数字化、人文、探讨
4	智慧图书馆	图书馆、建设、智慧、治理、阅读	数据治理	数据、本文、治理、提出、保护
5	数据治理	数据、数字、开放、人文、治理	数据共享	本文、档案馆、开发、数据、共享
6	影响因素	影响、因素、感知、档案学、意愿	数字记忆	本文、记忆、认知、情绪、老年人
7	用户画像	画像、分类、学习、提出、风险	信息素养	本文、教育、素养、提出、探讨
8	电子文件	电子、文件、文书、证据、公众	应急事件	本文、应急、事件、学术、文书
9	信息素养	素养、教育、个人信息、保护、评论	影响因素	影响、因素、本文、社交、发现

由表 4 可知，ChatGPT 生成与学者撰写摘要的主题分布较为一致，如“智慧图书馆”、“数据治理”、“影响因素”、“信息素养”等。其中的差异主题主要体现在“学科体系”、“应急事件”、“数字人文”。表明二者在这些主题上的行文风格差异较大。

3.2.3 ROUGE 评测

为了定量评测 ChatGPT 生成与学者撰写的摘要文本之间的相似度，本文采用了最常用

于评价自动文本摘要效果的ROUGE-1、ROUGE-2、ROUGE-L，此外还增加了余弦相似度来检测二者之间的相似程度。ROUGE主要统计二者之间重叠的基本单元数目，而余弦相似度主要测量二者在整体方向上的相似性，结果如表8所示。

表8 ROUGE与余弦相似度评测结果

评测标准	学科名称			
	图书馆学	情报学	档案学	整体
ROUGE-1	29.69%	**31.06%**	28.50%	29.75%
ROUGE-2	7.92%	**8.49%**	7.49%	7.97%
ROUGE-3	22.04%	**22.10%**	21.66%	21.93%
余弦相似度	**73.83%**	73.30%	70.23%	72.45%

受当前ChatGPT输入字符的限制，且ChatGPT本身是作为一种生成式人工智能工具，其主要功能是由用户输入来生成回复内容。因此，在ROUGE评测时，暂无法与目前的基准算法进行对比，原因在于相关的基准算法是根据全文来生成对应的摘要。通过与当前主流基准算法在公开评测数据集上的表现对比可知[22,23]，尽管ChatGPT生成的摘要评测分数还偏低，但是其余弦相似度较高。这表明二者生成的内容主要存在"形似"现象，即表明看起来很接近，但是重叠单元数据较低。在三个二级学科领域中，情报学领域二者的相似度更高，而档案学领域二者的相似度更低，说明与情报学相比，档案学领域ChatGPT生成的内容与学者撰写的摘要内容差异更大。

4 结论

本文以信息资源管理领域的三个二级学科近年来高被引论文为研究对象，在获取论文标题的基础上，通过调用ChatGPT接口设计Prompt提问来获取AI生成的论文摘要。采用9种机器学习及深度学习算法对二者进行分类识别，并从文本内容的多个角度分析了二者的差异。

在分类识别上，主流的机器学习或深度学习分类模型可以有效识别摘要文本是ChatGPT生成还是学者撰写。在所选的9种分类模型中，两种深度学习模型ERNIE和BERT的分类效果最好，在机器学习算法中，除NB和KNN以外，其余5种机器学习算法的整体F1-Score均超过了90%。从二级学科来看，各分类算法在档案学领域的摘要分类上F1-Score较图情领域更低。

在文本特征分析上，从字的角度来看，ChatGPT生成的摘要平均长度比学者撰写的更长，二者的数据分散程度相当，峰度差距较大。具体到三个二级学科来看，在平均值指标上，情报学差距最小，档案学差距最大；在均方差指标上，情报学差距最小，档案学差距最大；在偏态指标上，情报学差距最小，图书馆学差距最大；在峰度指标上，情报学差距最小，图书馆学差距最大。从词角度来看，二者的整体关键词较为接近，但ChatGPT生成摘要文本往往伴随着"本文"、"提出"、"最后"这样的过渡性词语出现，从三个二级学科来看，情报学领域二者的关键词更为接近。从句的角度来看，ChatGPT生成摘要是学者撰写摘要的文本句子数量的近2倍，均方差也更高，二者的偏态相当，学者撰写摘要句子数量的峰度更大。具体到三个二级学科来说，在平均值指标上，档案学差距最大，图情领域差距更小；在均方差指标上，图书馆学差距最小，档案学差距最大；在偏态指标上，情报学差距最小，档案学差距最大；在峰度指标上，情报学差距最小，图书馆学差距最大。

在主题模型分析方面，ChatGPT生成与学者撰写摘要文本的主题分布较为一致，主要在"学科体系"、"应急事件"、"数字人文"方面有较大差异。在ROUGE评测方面，当前ChatGPT生成摘要的评测分数在ROUGE-1、ROUGE-2、ROUGE-L三个指标上均较低，但是其余弦相似度较高，表明ChatGPT生成的摘要文本存在"形似"而不"神似"的现象。在三个二级学科领域上，情报学的评分最高，档案学的评分最低，表明档案学领域ChatGPT

生成摘要文本与学者撰写的摘要文本差异更大。

本文还存在一些不足，受当前 ChatGPT 调用接口的影响，本文采用了基于 GPT-3.5 的 ChatGPT 来生成对应的摘要文本，而未采用最新的 GPT-4 模型。且当前的 ChatGPT 在输入字符数上有诸多限制，未来将从引言、正文、结论等部分进行研究，以期更为全面地分析二者之间的差异。此外，本文目前仅以信息资源管理领域的中文学术论文为研究对象，在未来的研究中将考虑对不同学科领域的论文进行对比分析。

参考文献：

[1] Wang F Y, Li J, Qin R, et al. ChatGPT for Computational Social Systems: From Conversational Applications to Human-Oriented Operating Systems[J]. IEEE Transactions on Computational Social Systems, 2023, 10(2): 414-425.
[2] Mondal S, Das S, Vrana V G. How to Bell the Cat? A Theoretical Review of Generative Artificial Intelligence towards Digital Disruption in All Walks of Life[J]. Technologies, 2023, 11(2): 44.
[3] Cheung K S. Real Estate Insights Unleashing the potential of ChatGPT in property valuation reports: the "Red Book" compliance Chain-of-thought (CoT) prompt engineering [J]. Journal of Property Investment & Finance, 2023, ahead-of-print(ahead-of-print).
[4] Productive Teaching Tool or Innovative Cheating?[EB/OL]. [2023-08-28]. https://study.com/resources/perceptions-of-ChatGPT-in-schools.
[5] Initial submission | Nature[EB/OL]. [2023-08-28]. https://www.nature.com/nature/for-authors/initial-submission.
[6] 图书情报工作 AI 政策声明[EB/OL]. [2023-08-28]. https://www.lis.ac.cn/CN/column/column27.shtml.
[7] Introducing ChatGPT[EB/OL]. [2023-03-30]. https://openai.com/blog/ChatGPT.
[8] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science [J]. Nature, 2023, 614(7947): 214-216.
[9] Budhwar P, Chowdhury S, Wood G, et al. Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT[J]. Human Resource Management Journal, 2023, 33(3): 606-659.
[10] 戴岭，胡姣，祝智庭. ChatGPT 赋能教育数字化转型的新方略[J]. 开放教育研究，2023，29(4): 41-48.
[11] 陆伟，刘家伟，马永强，等. ChatGPT 为代表的大模型对信息资源管理的影响[J]. 图书情报知识，2023，40(2): 6-9+70.
[12] 张智雄，于改红，刘熠，等. ChatGPT 对文献情报工作的影响[J]. 数据分析与知识发现，2023，7(3): 36-42.
[13] Lund B D, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries?[J]. Library Hi Tech News, 2023, 40(3): 26-29.
[14] 曹树金，曹茹烨. 从 ChatGPT 看生成式 AI 对情报学研究与实践的影响[J]. 现代情报，2023，43(4): 3-10.
[15] 周文欢. ChatGPT 在档案领域应用和意义[J]. 中国档案，2023(3): 62-63.
[16] Zheng H, Zhan H. ChatGPT in Scientific Writing: A Cautionary Tale[J]. The American Journal of Medicine, 2023, 136(8): 725-726.e6.
[17] ScientistSeesSquirrel. How to use ChatGPT in scientific writing[EB/OL]. (2023-06-20) [2023-08-11]. https://scientistseessquirrel.wordpress.com/2023/06/20/how-to-use-ChatGPT-i

n-scientific-writing/.
[18] 张华平，李林翰，李春锦. ChatGPT 中文性能测评与风险应对[J]. 数据分析与知识发现，2023，7(3)：16-25.
[19] 鲍彤，章成志. ChatGPT 中文信息抽取能力测评——以三种典型的抽取任务为例[J]. 数据分析与知识发现：1-16.
[20] 施亦龙，许鑫. ChatGPT 机器回答与知乎人工回答的比较[J]. 图书馆论坛：1-10.
[21] Nigh M. ChatGPT3 Prompt Engineering[CP/OL]. (2023-08-12)[2023-08-12]. https://github.com/mattnigh/ChatGPT3-Free-Prompt-List.
[22] 王红斌，金子铃，毛存礼. 结合层级注意力的抽取式新闻文本自动摘要[J]. 计算机科学与探索，2022，16(4)：877-887.
[23] 赵江江，王洋，许楹楹，等. 基于知识蒸馏的抽取式自动摘要模型[J]. 计算机科学，2023，50(S1)：214-220.

作者贡献说明：张强：设计研究方案，实验分析与处理，论文初稿的撰写与修改；王潇冉：进行实验，实验结果分析；高颖：论文修改及最终版本修订；周洪：提出论文修改意见。

Comparative Study on ChatGPT Generation and Scholars Writing of Literature Abstracts: Taking the Field of Information Resource Management as an Example

Zhang Qiang[1], Wang Xiaoran[2], Gao Ying[1], Zhou Hong[3,4]
(1.School of Information Management, Central China Normal University, Wuhan 430079; School of Computer and Information, Anhui Polytechnic University, Wuhu 241000; 3. Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100191, China; 4. Wuhan Library and Intelligence Center of Chinese Academy of Science, Wuhan 430071)

Abstract： [Purpose/Significance] Explore the similarities and differences between ChatGPT generation and Chinese paper abstracts written by scholars, and analyze the differences in content characteristics between the two, providing reference for AI generated academic paper detection and related research. [Method/Process] Firstly, taking the field of information resource management as an example, we extracted 500 highly cited papers from library science, information science, and archival science in the past three years. Based on the obtained paper titles, we used the Prompt method to apply the ChatGPT tool to generate corresponding abstract texts and construct a dataset; Secondly, 9 machine learning and deep learning algorithms were used to classify and detect abstract texts generated by ChatGPT and written by scholars; Finally, analyze the similarities and differences between the two from multiple perspectives, including text features, topic models, and ROUGE evaluation, in order to reveal the similarities and differences between the two. [Result/Conclusion] Mainstream machine learning and deep learning algorithms trained on datasets can effectively distinguish whether abstracts are generated by AI or written by scholars, with BERT and ERNIE performing the best, while RF and Xgboost perform the best among machine learning algorithms. The number of abstract characters and sentences generated by ChatGPT is higher than that written by scholars, and the keywords are mostly template based transitional words; The themes of the two texts are mostly the same, but there are differences in themes such as "disciplinary system" and "digital humanities"; The quantitative analysis of

ROUGE and cosine similarity indicates that the abstracts generated by ChatGPT have a significant "resemblance" rather than a "resemblance" to the abstract texts written by scholars.
Keywords: ChatGPT Text classification Text features Paper abstract